Mapping and predicting research trends in international journal publications using graph and topic modeling

Asep Herman Suyanto[1], Taufik Djatna[2], Sony Hartono Wijaya[3]
[1]Postgraduate Program in Department of Computer Science, Faculty of Mathematics and Natural Sciences, Bogor Agricultural Institute (IPB) University, Bogor, Indonesia
[2]Department of Agro-Industrial Technology, Faculty of Agricultural Technology, Bogor Agricultural Institute (IPB) University, Bogor, Indonesia
[3]Department of Computer Science, Faculty of Mathematics and Natural Sciences, Bogor Agricultural Institute (IPB) University, Bogor, Indonesia

Article Info

Article history:

Received Jan 2, 2023
Revised Jan 15, 2023
Accepted Jan 25, 2023

Keywords:

Graph theory
International journal
Predicting trend
Research mapping
Topic modeling

ABSTRACT

Researchers and journal managers need summary information, such as research maps and trends. Topic and words-based document content analysis alternative to science mapping and trend prediction based on bibliographic analysis. The data are a collection of journal articles/proceeding documents and metadata for 2011-2020 published by the International Journal of Electrical and Computer Engineering (IJECE). A combination of several techniques and methods is used, such as text mining, topic modeling, cosine similarity, network analysis, graph theory, and seasonal autoregressive integrated moving average (SARIMA). This research has produced research topics, mapping, trend prediction, and visualization according to its objectives. The results of the topic coherence test obtained the optimal number of topics, as many as 6. The results of the topics were evaluated by experts to be given labels and areas of focus. On the research map for each topic, information is found on trending research, the most popular research, research that is central to the research group, and critical research on the development of the group path. It also identifies the type of breakthrough, incremental, and research gap. Predictions of research trends obtained are based on topics and words that describe the development of research. Visualization is descriptive and predictive.

Corresponding Author:

Taufik Djatna
Department of Agro-Industrial Technology, Faculty of Agricultural Technology
Bogor Agricultural Institute (IPB) University
Fateta Building 2nd Floor, Dramaga Campus of IPB Bogor, Indonesia
Email: taufikdjatna@apps.ipb.ac.id

1. INTRODUCTION

The development of scientific publications is currently increasing drastically, as shown by the indexing journal. A large number of scientific publications and their rapid growth mean that the scope of research is now enormous and its substance so complex that personal experience is no longer sufficient for quantitative analysis, understanding trends, or making decisions without the help of systems [1]. With so many scientific publications, summary information such as maps and research trends is indispensable.

Scientific publications consist of scientific articles, review articles, books, sections, book reviews, theses, reports, and patents [2]. Meanwhile, only patents and research articles can be indicators of research mapping because they are related to the development of science and research [3]. Because of that, research articles can be a source of data for mapping and predicting research trends.

International journals are scientific journals written in accordance with scientific principles and scientific ethics. It has an international standard serial number (ISSN) and an online version of the publication. Experts in their field of specialization from at least four countries make up the editorial board. Arabic, Chinese, French, English, Russian and Spanish are the official United Nations (UN) languages for journal writing. At least two countries are represented in scientific articles published in one publication number. They are published by a world-renowned professional association or a credible university or publisher. They are Indexed by an international rating or a well-known international database. The review process is carried out correctly and adequately [4].

Science/research maps are a set of tools and approaches for mapping the structure and evolution of knowledge in an area or field utilizing maps as a visual communication metaphor. A science mapping study's scope can include a scientific profession, a field of research, or specific issue areas related to research themes. Research maps comprise a collection of scientific papers that have been evaluated with computer techniques and visualized to show trends that may be understood using a scientific theory of change [5], [6].

Bibliographic analysis techniques like citation/co-citation, co-authorship, co-occurrence/co-word, and bibliographic coupling are often utilized in science mapping. The citation/co-citation analysis is used to analyze the relationship between cited documents and understand the development of fundamental topics in the research field. In order to understand the periodicity or current development of the research object, bibliographic coupling analyses are based on the relationship and similarity between the same references. Co-word analysis the appearance of keywords and identifies the relationship/interaction between topics and emerging trends by focusing on the content of the document itself. Co-authoring analysis examines the social interaction or relationship between the author and his or her affiliation and the implications for the development of the research field [7].

Citation-based analysis and bibliographic coupling have some drawbacks. First, the analysis focuses on the linkage of the article document because it is challenging to know the linkage as a whole. Second, there is limited information obtained because the built relationships are only limited to quoting relationships. Third, the internal linkages between articles cannot be explored to reflect the actual relationship. Fourth, it takes much time due to the deep search process [8]. Fifth, a high number of citations does not necessarily indicate quality, perhaps citing their own output or citing output from the journals in which they publish [9]. Document content analysis overcomes these limitations, which focus more on managing and analyzing the knowledge contained in documents and reflecting on the actual relationship, as well as knowing the overall relationship between documents [8], [10], [11].

In the text mining comparison test on scientific articles, it was found that the full-text performance improvement was better than the title/abstract; therefore, all full-text information is indeed more valuable and needed [12], [13]. Analyzing the entire text using topic modeling can enable discovering the thematic/topical structure underlying the paper, studying how topics evolve, and predicting future trends [9]. One method is latent dirichlet allocation (LDA) which has the advantage of interpreting semantically, does not require training data, and can handle long documents. LDA generates topics based on word contribution probabilities, making it simple for analysts to understand each topic's category. The probability of topic contribution for each document is also more precise because it makes the grouping process easier [14], [15]. Several comparative research of topic modeling techniques shows that LDA is better (state of the art) than other topic modeling techniques, especially for long texts [15]–[17].

The two types of research document analysis are network-based and keyword-based. Internal information cannot be evaluated in a network-based system. On the contrary, the relationship between documents in the keyword-based analysis does not exist [18]. The approach to combining this method is text mining, which extracts keywords and composes document vectors to see similarities between documents [8]. Cosine similarity can compare documents by measuring the frequency of certain words (vectors) in the document [19]. The linkages between documents are visualized in a graph form, allowing network analysis to be carried out to obtain information about activities on the research map [8]. Various measures are used for network analysis, such as centrality, representing the strength of an edge and node in the network based on their role and position [20].

This type of research/innovation is usually divided into two major parts, breakthrough and incremental. The research included in the breakthrough group is research without prior research; on the other hand, the research included in the incremental group is a development/continuation research from previous research in a research group [21]. The definition and scope adopted by this research are to distinguish preliminary research and further research so that the series of research developments are known.

To detect research gaps in existing research, a researcher must collect related research works, review, code information, and narratively present them using the traditional method of finding research ideas. To get to the gray areas, critically examine confidential information and trends in previous research, create visual data representations, and summarize information using appropriate metrics. The process is not easy; therefore, one can use graph theory principles [22]. It is common to find research gaps on research maps, defined as topics or areas with missing or insufficient information, limiting reviewers' ability to reach conclusions. Research gaps can be developed further, such as through stakeholder involvement in prioritizing research needs. Identifying a clear and explicit research gap is essential in developing a research agenda [23].

Research trends can be predicted using subjective judgments, but expert results can be biased so quantitative analyzes are used, such as bibliometrics, scientometrics, or informetrics. However, it has limitations because the prediction of topic-based research trends on information throughout the document's contents can be an alternative approach [9]. Mapping and research trends must have time series used to measure at various moments of time intervals that describe and identify trends, periodicals, and predictions [24]. Time-series data having a seasonal trend, modeling and forecasting can be applied with seasonal autoregressive integrated moving average (SARIMA) [25]. In the amount of research that tested seasonal data, it was found that the SARIMA model had better performance than several models, such as autoregressive–moving-average (ARMA), autoregressive moving average exogenous (ARMAX), autoregressive integrated moving average (ARIMA), convolutional neural network (CNN), and long short-term memory (LSTM) [26]–[28].

This research is inspired by previous research, although it differs in the implementation of the case studies. First, Yoon and Park [8] use text mining and network analysis to create patent networks that show high-tech trends and identify new product development avenues. Second, Hakim *et al.* [21] make a technology map in Indonesian-language journals using text mining and network analysis. Third, Abuhay *et al.* [9] predict trends in scientific research topics in proceedings using non-negative matrix factorization (NMF) topic modeling and ARIMA prediction methods. Fourth, Lafia *et al.* [5] research mapping topics at various levels on the title, abstract, and keyword data of research documents using LDA and non-negative matrix factorization (NMF) topic modeling comparisons. Then design a science map that reveals the thematic structure of interdisciplinary research using t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) visualizations.

This research aims to produce research topics, research mapping, prediction of research trends, and visualization. Using topic and word-based content analysis as an alternative to research mapping and prediction using bibliographic analysis. The innovation/novelty claims of this research from before are analyzing not only the title/abstract but the entire text of the document by combining techniques and methods, including text mining, topic modeling LDA, cosine similarity, network analysis, graph theory, and seasonal autoregressive integrated moving average (SARIMA).

2. METHOD

2.1. Research stages

This research stage adopted the knowledge discovery in database (KDD) process [29]. The method was adopted with several adjustments, as shown in Figure 1, starting from data collection, data integration, data cleaning, pre-processing, transformation, and feature selection. Then in parallel, calculate the similarity between documents using cosine similarity and search for topics using latent dirichlet allocation (LDA). The results are combined to create a research map by calculating the network analysis on each topic using degree, eigenvector centrality, closeness centrality, and betweenness centrality. Then identify the type of research on each topic into breakthrough research or incremental research. The trend development results from the relationship between the year of publication and the number of publications on each topic. The results of term frequency-inverse document frequency (TF-IDF) and n-grams (bigrams and trigrams) related to the year of publication are used to determine the evolution of word/based research trends. Meanwhile, to get future trend predictions using the seasonal autoregressive integrated moving average (SARIMA) method. The final result of the research map is visualized into a graph, and the prediction of research trends is visualized in a time series.

2.2. Data collection

The data are a collection of journal articles/proceeding documents and metadata for 2011-2020 published by the International Journal of Electrical and Computer Engineering (IJECE). Documents in the form of surveys and review papers were not used in this research. The data collected and used were 3,091 documents from 56 volumes. Metadata and portable document format (pdf) documents were obtained using web scraping from the online journal ijece.iaescore.com. The pdf document is converted into plaintext, and the text cleaning process is carried out, then integrated with metadata selected the required attributes, then saved to database/Excel. Figure 2 shows the number of articles collected from 2011-2020.

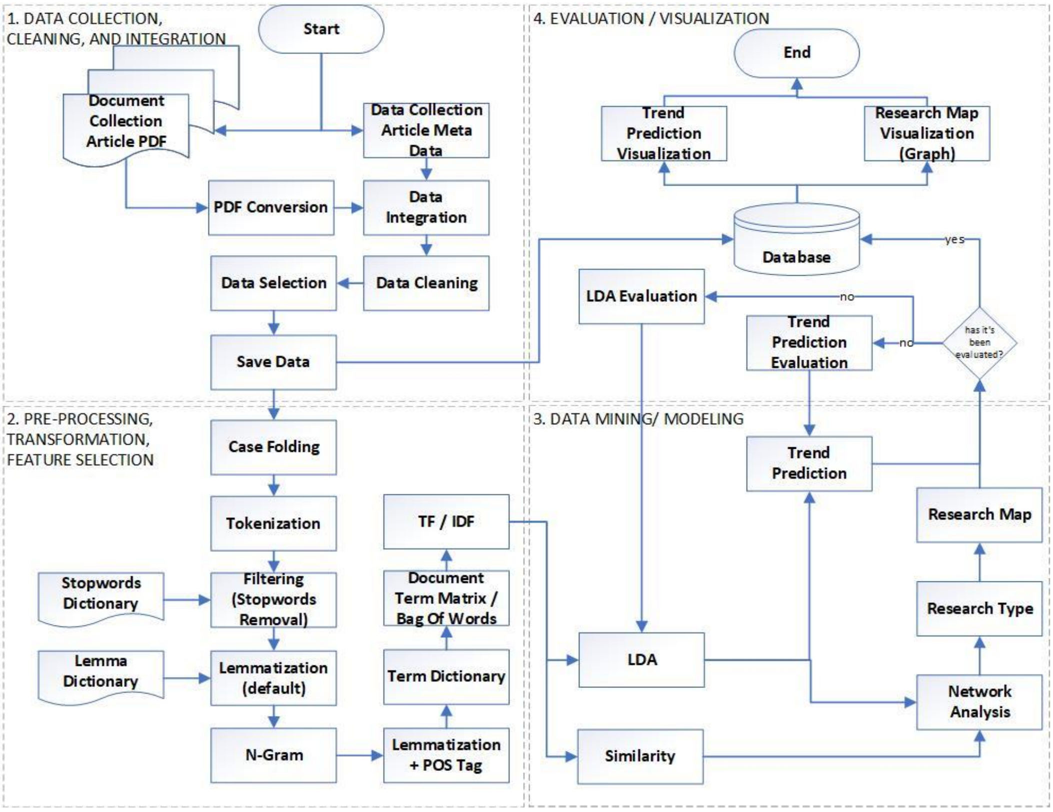


Figure 1. Research stages

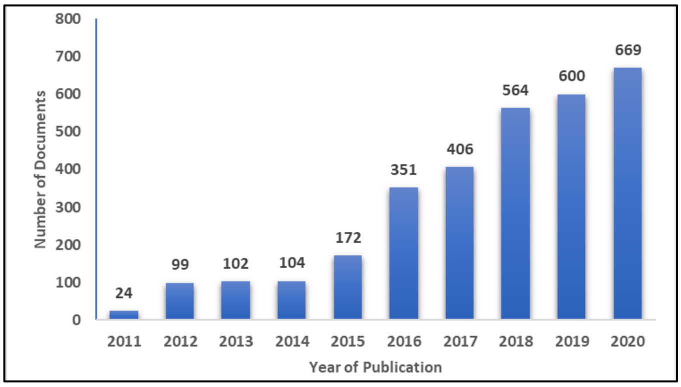


Figure 2. Number of article documents each year

2.3. Data preprocessing

Case folding converts text to lowercase and removes all characters without letters a-z, to make it easier to compare text and more consistent data during pre-processing. Tokenization is tokenization to cut text into several parts called tokens. Filtering is the stage of taking essential words from the token results and discarding stopwords. Stopwords need to be removed because they make the composition of the text larger however do not contribute meaning to the context or the text of the document. Eliminating these stopwords can reduce index size, processing time and reduce noise. Lemmatization is the computational method of determining a word's lemma (dictionary word) based on its intended meaning. To get consistent lemmatization results, it is necessary to use part of speech taggers that classify words as nouns, and verbs [30].

N-gram is a set of n words extracted or clipped from a text commonly used in language modeling. Size one is a unigram, size two is a bigram, and size three is a trigram, and so on. The N-gram uses an efficient skip-gram model to explore high-quality distributed vector representations that capture numerous exact syntactic and semantic word associations [31].

Text transformation to get a representation of the document represented by the words (features) that appear. This research found the dictionary of terms which is a collection of unique indexed words. The bag-of-words (BoW) is a program that counts the number of times a word appears in a document while ignoring the order of the words [32]. In general, finding different themes or topics using a minimum word frequency document depends on the corpus's size. Document frequencies of 10, 20, 50, 100 are a good starting point [33].

The process of picking a subset of significant features for modeling is known as feature selection. A global dictionary can be created from all papers in the collection to create a local dictionary. Unique feature selection methods will pick a subset of words that appear to have the most predictive potential [10]. Term frequency-inverse document frequency (TF-IDF) feature selection is a statistical metric that describes the value of words in a document in a document collection. TF-IDF uses the inverse proportion of the documents containing that word to calculate the relative frequency of each unique word in a single text. The primary notion is that phrases that frequently appear in a large number of texts are given less weight than words that frequently appear in a single document [34]–[36].

2.4. Topic modeling

Latent dirichlet allocation (LDA) is a modeling subject on document collections that uses a generative probability model to make document processing more efficient and explicitly represent the content [37]. LDA uses (1) [38], where α is the parameter of proportion, η is topic parameter, β_k is the distribution of words to the topic, θ_d is the topic distribution for document d, K is a number of topics, N is a number of words, D is a number of documents, $Z_{d,n}$ is the topic for the n-th word in document d, $W_{d,n}$ is the n-th word in document d, which is an element of the corpus. Figure 3 shows a graphic of the LDA model. Evaluation of LDA uses the topic coherence C_V (coherence value) method to find the optimum topics. C_V based on sliding window, top-word segmentation, and indirect confirmation measures employing normalized pointwise mutual information (NPMI) and cosine similarity merging [39].

$$p(\beta,\theta,z,w|\alpha,\eta) = \prod_{i=1}^{K} p(\beta_i|\,\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \left(\prod_{n=1}^{N} p\left(Z_{d,n}|\theta_d\right) p(W_{d,n}|\beta_{i,k}, Z_{d,n})\right) \quad (1)$$

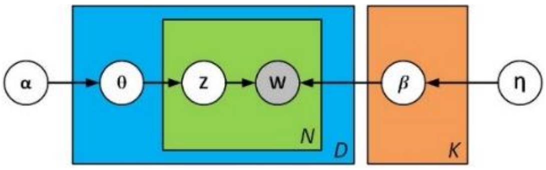


Figure 3. The graphical model for latent dirichlet allocation [38]

2.5. Cosine similarity

Documents can be represented by any specific word frequency attributes in the document called vectors. Cosine similarity to find similarities between documents, the greater the value of the equation, the more similar the two documents. The resulting similarity is measured on a scale of 0 to 1, with values in the middle indicating similarity. Cosine similarity is shown in (2). A_i is the i-document vector in the A-document set, B $_i$is the i-document vector in the B-document set, and n is the number of documents evaluated [40].

$$similarity = cos(\theta) = \frac{A.B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (2)$$

2.6. Graph and network analysis

Graph theory is the study of graphs represented by a collection of edges and nodes based on a mathematical structure [41]. Documents that have a relationship are modeled into the concept of a graph. The graph is a set G={V, E}, where V is the set of nodes V={v_1, v_2, v_3,..., v_i} and E is the set of edges E={v_1v_2, v_1v_3,..., v_iv_j}. In this research, V denotes the analyzed documents, for E is the relationship between two documents assessed by their association values [21]. Because relationships are visualized as graphs, network analysis can display crucial information from research graphs and simulate research growth activities. This method has the advantage of simplifying and identifying internal linkages between documents [8].

The value of disciplinary or community collaboration across research domains is measured through network analysis, which quantifies the importance of a node (document) in a network [42]. Measurements often used in research maps are degree, eigenvector centrality, closeness centrality, and betweenness centrality—an illustration of network centrality, as shown in Figure 4.

Degree shows the number of correlations with other research. Research with a higher degree value indicates that the research is central/popular from other research and is becoming a trend. The degree can be calculated by (3) and illustration shown in Figure 4(a), where A_{ij}=1 if there is an edge between i and j, otherwise, it will be 0, and n is the number of nodes in the graph [43].

$$k_i = \sum_{j=1}^{n} A_{ij} \tag{3}$$

Closeness centrality shows that the research covers many aspects of a research topic and strongly connects to the research topic. Closeness centrality can be calculated by (4) and illustration shown in Figure 4(b), where $\mathrm{d}(v, y)$ is the distance between vertices *v* and *y*, and k is the number of nodes connected to *v* [44].

$$\mathrm{C}_c(v) = \frac{1}{k} \sum_y \frac{1}{\mathrm{d}(v,y)} \tag{4}$$

Betweenness centrality higher level has a higher level of interest in research development. Betweenness centrality can be calculated by (5) and illustration shown in Figure 4(c), where the sum is taken over all distinct pairs s and t, σ_{st} is the number of shortest paths from s to t, and $\sigma_{st}(v)$ is the number of these paths that pass through v, and set this ratio to be 0 if there are no paths from s to t [44].

$$\mathrm{C}_b(v) = \frac{1}{2} \sum_{s \neq t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{5}$$

Eigenvector centrality, a higher one, indicates the influence/importance of the research because it is linked more by other important/popular research. Eigenvector centrality can be calculated by (6) and illustration shown in Figure 4(d), where A_{ij}=1 if there is an edge between i and j; otherwise, it will be 0. $\mathcal{X}_j$ is the eigenvector centrality of node j and λ constant value [43].

$$\mathcal{X}_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij}\, \mathcal{X}_j \tag{6}$$

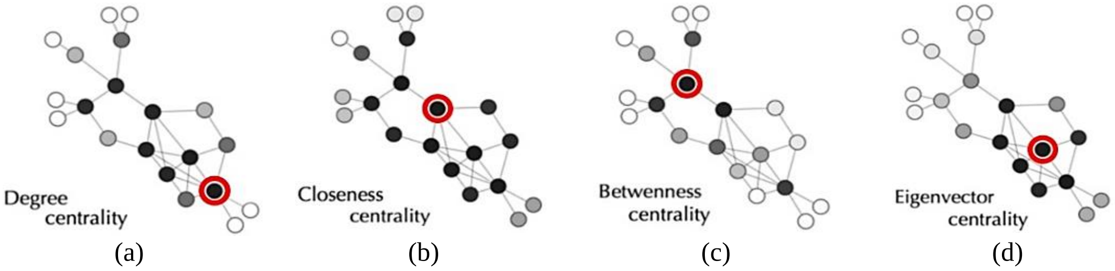


Figure 4. Illustration of network centrality [45] (a) degree centrality, (b) closeness centrality, (c) betweenness centrality, and (d) eigenvector centrality

2.7. Research type

Documents on each topic are identified as breakthrough research or incremental research. Breakthrough research is without prior research, while incremental research is developed from previous research on a topic. The type of research T(*vi*) is defined by (7), where I_{ij} is a binary value that has a value of 1 if research is a breakthrough from research j and 0 if research is incremental from research j [21].

$$T(vi) = \prod_{j=1}^{J} I_{ij} \tag{7}$$

2.8. SARIMA

Each topic and word can be linked to a time series to obtain developmental information and predict research trends. Meanwhile, to get future trend predictions using SARIMA forecasting method. SARIMA forecasting modeling is denoted by SARIMA(p,d,q)(P,D,Q)$_s$. The (p,d,q) order of the non-seasonal component of the model for the autoregressive (AR) parameters, differences, moving-average (MA) parameters. The (P,D,Q) order of the seasonal component of the model for the AR parameters, differences, MA parameters, and s is the periodicity of the seasons [25].

The performance of the best prediction model is determined by examining the Akaike's information criterion (AIC) value, which is based on the parsimony principle that the greater the number of model parameters, the greater the complexity, and a solution must be found that allows for a balance of adaptability and complexity [25]. A good model is a model with the smallest AIC value. AIC can be calculated as in (8) [46]. K is the number of parameters in the model, and $log{-}likelihood$ is a measure of the accuracy of the model where the higher the value, the more precise.

$$\text{AIC} = -2\,(log - \text{likelihood}) + 2\text{K} \quad (8)$$

The evaluation of the forecasting error used in this research is the root mean square error (RMSE). The model gets better when the resulting RMSE value is close to zero, which means the closer the predicted and observed values are. RSME can be calculated as in (9) [47], where y is the observed value, ŷ is the value of the predicted result, i is the order of the data, and n is the number of data.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (9)$$

3. RESULTS AND DISCUSSION

3.1. Data preprocessing

The pdf text used is the title, abstract, keywords, introduction, method, results and discussion, conclusion, and suggestions. After data cleaning, the text results from the pdf and metadata are integrated and stored. The data obtained consists of attributes identification (ID), volume, year, file, title, author, uniform resource locator (URL), page, and text. In the case folding stage, all text becomes lowercase, deletes all characters that do not include letters a-z, and ignores words less than three letters. Tokenization refers to the process of breaking down a text into tokens (words). The filtering stage removes the stopword used and common words found in journals and documents. Lemmatization before n-gram without parts-of-speech (POS), while after n-gram apply with POS because the results of n-gram are still understood. N-grams used are unigrams, bigrams, and trigrams, to maintain semantics and define phrases. The parameter used ignores all words and n-grams with a number lower than 20 in the entire corpus. The threshold for bigram 20 and trigram 20, if the threshold value is higher, then the resulting phrase will be less. The transformation stage creates a dictionary with the results of 5,114 vocabularies, then creates a BoW. The threshold is used to remove all words that appear less than 30 times in all documents to eliminate spam/false words. The maximum threshold used is 50% to appear throughout the document to eliminate overly common words. The following processing is the feature selection process using TF-IDF to give weighting words in the document collection.

3.2. Topic modeling

A grid search on LDA is needed to find the optimal number of topics using topic coherence calculations with C_V. If the test results are more significant, the better interpretation results will be. In the test results from the number of topics 2 to 20, the best value is produced with a coherence score of 0.4600 on the number of topics 6, as shown in Figure 5.

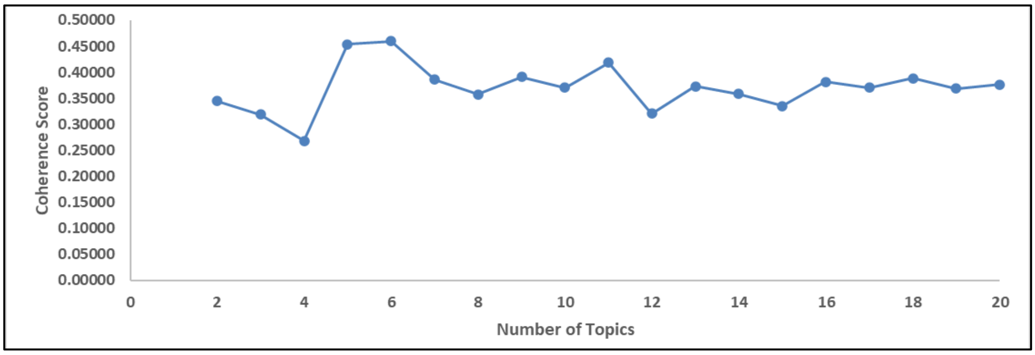


Figure 5. Topic coherence test results

The corpus, dictionary, number of topics, and number of words per subject are all parameters that must be included in the LDA model procedure. The extraction results in the form of words are evaluated by experts to determine the topic label and the focus area of the topic, as shown in Table 1. LDA results also get the distribution of topics in the documents used to identify topics in each document. These results are used to predict research trends and research maps at the next stage. Phrases from the pre-processed LDA result set were selected and used as sub-topics with non-general terms.

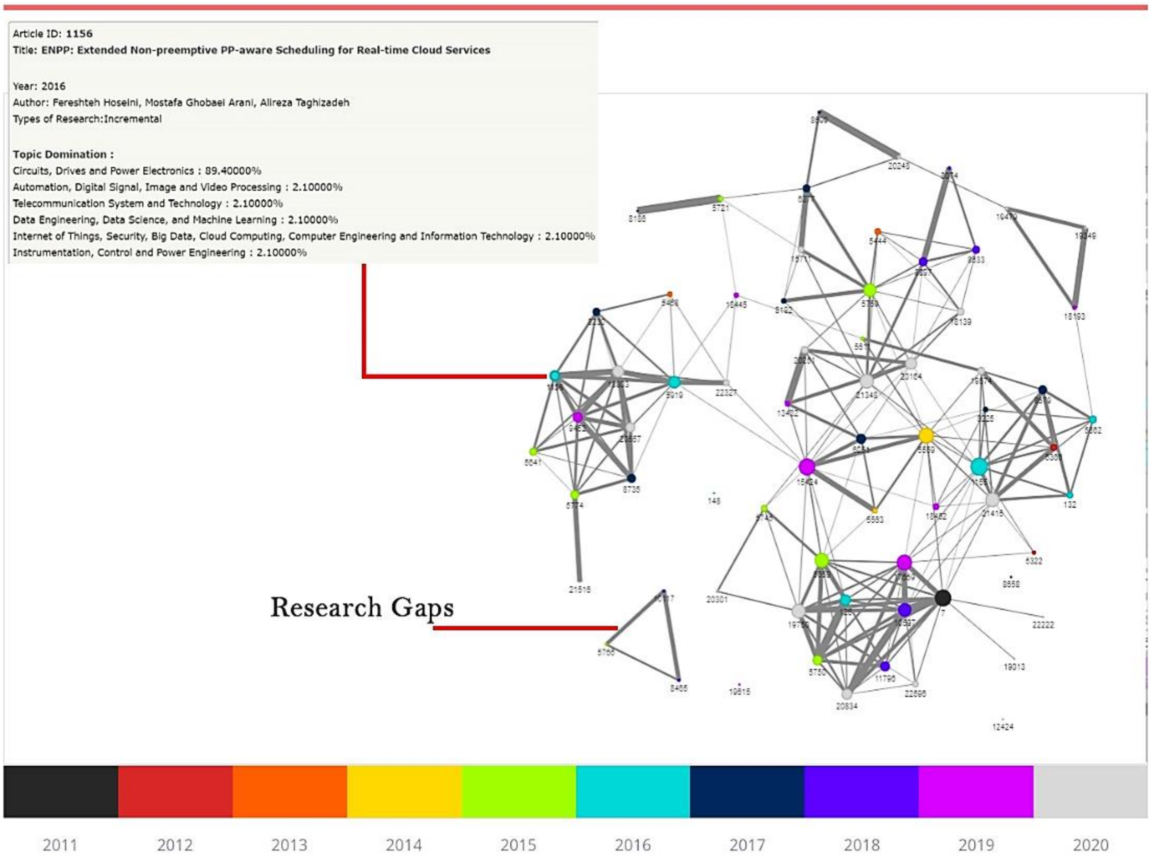
Table 1. LDA model results

Topic ID	Topic labels	Focus area
1	Automation, Digital Signal, Image, and Video Processing	Electrical & Electronics Engineering
2	Telecommunication System and Technology	Electrical & Electronics Engineering
3	Data Engineering, Data Science, and Machine Learning	Computer Science
4	Circuits, Drives, and Power Electronics	Electrical & Electronics Engineering
5	Internet of Things, Security, Big Data, Cloud Computing, Computer Engineering, and Information Technology	Computer Science
6	Instrumentation, Control, and Power Engineering	Electrical & Electronics Engineering

3.3. Research mapping

The similarity calculation between documents uses cosine similarity, which produces a value between 0 to 1. This value is used as the basis for determining the relationship between one research and another. By applying the preliminary threshold value of 0.1, it was obtained as many as 288,361 similarity comparisons. Each topic implements a different cut-off value in the similarity calculation results. The visualization of the research map is still clearly observed and forms a more meaningful relationship between documents.

The research map in Figure 6 is an example of a visualization of a research map for the topic circuits, drives, and power electronics during 2011-2020. On the research map, some nodes show research documents and edges that show research links. The larger the node size, the greater the relationship with much other research, and vice versa. The thicker the edge, the stronger the similarities between the two research, and vice versa. Clicking on the node will display detailed information, the type of research, and the distribution of topics in the research. The colour indicates the year of research which helps identify the type of breakthrough or incremental.



Article ID: **1156**
Title: ENPP: Extended Non-preemptive PP-aware Scheduling for Real-time Cloud Services

Year: 2016
Author: Fereshteh Hoseini, Mostafa Ghobaei Arani, Alireza Taghizadeh
Types of Research:Incremental

Topic Domination :
Circuits, Drives and Power Electronics : 89.40000%
Automation, Digital Signal, Image and Video Processing : 2.10000%
Telecommunication System and Technology : 2.10000%
Data Engineering, Data Science, and Machine Learning : 2.10000%
Internet of Things, Security, Big Data, Cloud Computing, Computer Engineering and Information Technology : 2.10000%
Instrumentation, Control and Power Engineering : 2.10000%

Figure 6. Research mapping visualization for topic circuits, drives, and power electronics

The research gap can be seen from the differences in previous research results, including concepts, theories, data, problems, and others. This can be a gap for further research, but it needs more in-depth research. The process is not an easy task; therefore, it is necessary to use graph theory. Research gaps on the research map for each topic are found in research that has no relation to any network because related information is missing. Research gaps can also be seen in networks that are separate from the dominant network. They can potentially become a research trend or novelty if the publication is in the latest year.

Table 2 shows the highest network centrality value for each topic. Calculating the similarity of documents on each topic has resulted in research linkages. Research with a higher degree value indicates the center of other research and becomes a trend (trend research). Research with a high value of closeness centrality shows that the research covers many aspects of a research topic and has a solid connection to the research topic (central research group). Research with a high-value betweenness centrality is critical in research development (research is essential in developing group pathways). Research with a higher eigenvector centrality value shows the importance of the research than other research that is considered essential. Documents on each topic are identified as breakthrough research (br) and incremental research (in). Breakthrough is research without prior research on a topic. Incremental research is the development/continuation of previous research.

Table 2. The highest network centrality value for each topic

Topic	Cut		Degree			Closeness centrality			Betweenness centrality			Eigenvector centrality	
ID	Off	ID	Value	Type	ID	Value	Type	ID	Value	Type	ID	Value	Type
1	0.3	12742	24	in	12742	0.1596	in	20337	0.0310	in	12742	0.2979	in
2	0.5	745	66	in	745	0.5910	in	6449	0.0231	in	745	0.1576	in
3	0.5	6268	40	in	6268	0.1374	in	6268	0.0500	in	6268	0.3495	in
4	0.1	1165	15	in	15424	0.4371	in	15424	0.2991	in	17659	0.3171	in
5	0.5	15066	23	in	17821	0.1510	in	16851	0.1329	in	15066	0.2922	in
6	0.5	16735	28	in	707	0.0761	in	14818	0.1196	in	11735	0.3277	in

3.4. Research trend prediction

Topics and word results on each topic can also be related to the time series to obtain information on developments and predictions of research trends. Before producing the final prediction model, testing is needed to find the most optimal parameters using a grid search. This approach can be faster on a computer than a manual analysis process while revealing surprising findings that may not be obvious and result in lower estimation errors. The range of grid search parameters for 7 variables in the SARIMA(p,d,q)(P,D,Q)$_s$ model is 0 to 2. The evaluation determines the performance of the best prediction model for SARIMA by looking at the smallest AIC value. Then the evaluation of forecasting errors used is RMSE. The model gets better when the resulting RMSE value is close to zero. The search results for evaluating the best predictive models for each research topic are shown in Table 3. Evaluate the best prediction model for words-based using the same model parameters, namely SARIMA(1, 1, 1)x(1, 1, 1)$_6$.

Table 3. Evaluation of research topic trend prediction model

Topic ID	Model	Log Likelihood	AIC	RSME
1	SARIMA(0, 1, 1)x(0, 1, 1)$_6$	-49.73	105.45	2.79
2	SARIMA(0, 1, 1)x(0, 1, 1)$_6$	-61.37	128.74	4.01
3	SARIMA(0, 1, 1)x(0, 1, 1)$_6$	-85.99	177.98	20.54
4	SARIMA(0, 1, 1)x(1, 1, 1)$_6$	-40.29	86.57	1.48
5	SARIMA(1, 1, 1)x(0, 1, 1)$_6$	-80.94	169.88	33.08
6	SARIMA(0, 1, 1)x(1, 1, 1)$_6$	-87.69	183.38	13.50

Predicting research topic trend development is identified in the document related to the year of publication, as shown in Figure 7. The graph shows trends (2011-2020) and predictions of topics for the next five years (2021-2025), which a trend of rising in topics 6 (instrumentation, control, and power engineering), 3 (data engineering, data science, and machine learning), and 5 (internet of things, security, big data, cloud computing, computer engineering, and information technology). In comparison, the rest did not develop too significantly because it had not become the focus of the journal manager.

Predicting research trends can also be seen in depth from the words related to the year of publication and the TF-IDF weighting of words in the document. Predictors were obtained from the LDA results' terms, which only took the dominant n-grams (bigrams and trigrams) phrases for each topic, as shown in Figure 8. Users can select words in WordCloud. The color in the WordCloud text shows the sub-topic on a specific topic, as shown in Figure 8(a). Provides an overview of trend developments (2011-2020) and predictions of topics for the next five years (2021-2025), as shown in Figure 8(b). Future predictions on several sub-topics that will continue to develop and become trends are wind turbine/speed/energy, deep learning, machine learning, fuzzy logic, neural network, proportional–integral–derivative (PID) controller, and active/reactive power. Journal managers can consider the results of developments and predictions of trends in making future policies.

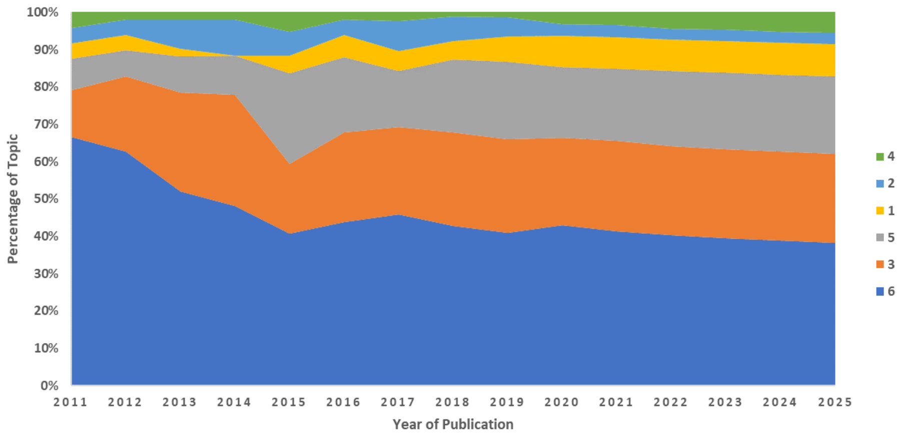
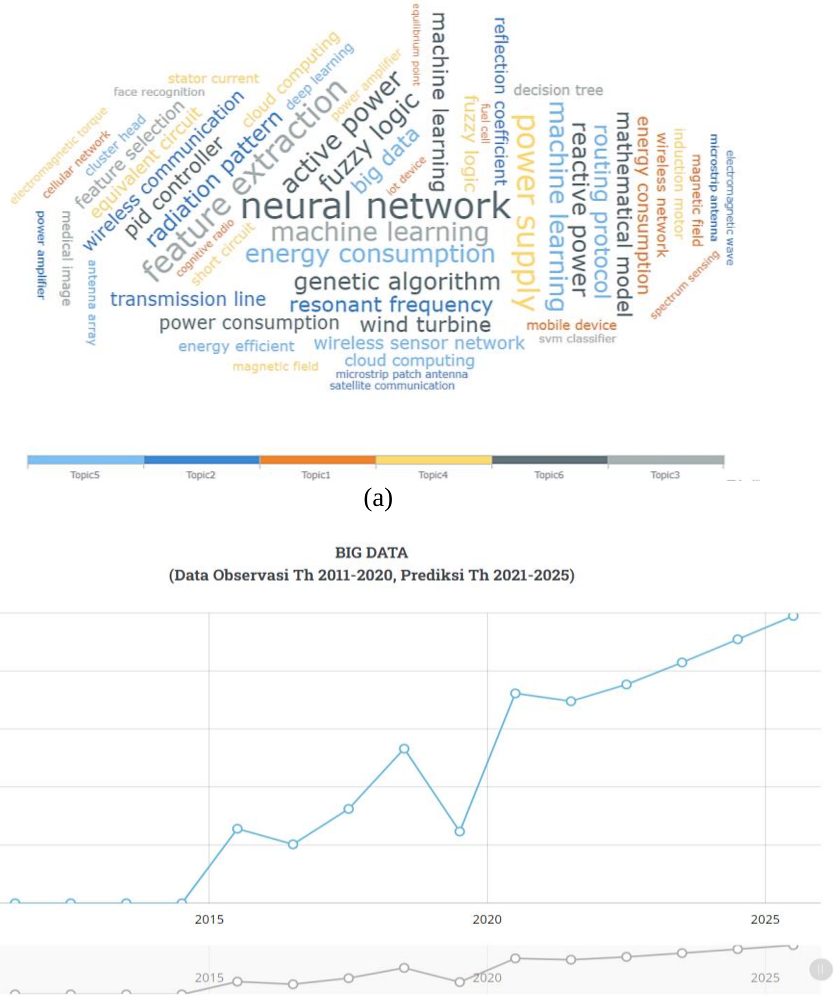


Figure 7. Topic-based prediction visualization



(a)

(b)

Figure 8. Words-based prediction visualization (a) wordcloud and (b) time series

This research has its advantages, disadvantages, and difficult parts. Superiority claims of this research use several combinations of techniques and methods, including text mining, LDA topic modeling, cosine similarity, network analysis, graph theory, and SARIMA for content analysis based on the topic and words as an alternative to research mapping and predicting research trends that have been done using bibliographical analysis. System visualization is web-based that can be accessed anytime, anywhere, and can be customized as needed. The drawbacks of the results of this research are still dependent on experts in giving labels to topics; besides, the results of the distribution of topics are too global. The system developed with Django is still limited to the front end, but backend processes are still using Jupiter. The difficulty of this research is the selection of the sequence of techniques at the pre-processing stage of the data so that some words/texts are sacrificed.

4. CONCLUSION

The results of the topic coherence test with C_V get an optimal number of topics, as many as 6. Experts analyze the findings of themes in the form of a collection of words to assign labels to the topic and area focus. Topic modeling also produces a distribution of topics in each document for grouping documents according to research topics. Topic identification in each document is used to map and predict research trends. In the results of the research map per topic, it is found that research information is trending, the most popular is the center of the research group, and is essential in the development of the group path. It also identifies the type of breakthrough, incremental, and research gap. This research map can be used to identify, evaluate, and obtain information on potential research areas. The predictions of research trends obtained are topic-based and word-based, which provide a comparative picture of the development of research trends in the next few years. Topic-based research trend predictions generate research topic trends every year and predictions for the next few years. Predicting research trends can also be seen in depth from the words related to the year of publication and the TF-IDF weighting of words in the document. Predictors were obtained from the LDA results' terms, which only took the dominant n-grams (bigrams and trigrams) phrases for each topic. Prediction of research trends can be used for consideration of research planning and development. Researchers/writers can plan on which topics will contribute. Journal managers can also evaluate research articles that will be accepted, whether they are relevant to current trends or complement research topics that are still empty, so that they become more focused on the purpose of the journal. Visualization mapping and predicting research trends are available in one system in the form of descriptive and predictive, making it easy to be accessed and understood so that they can be utilized and explored according to their respective needs. Suggestions for further development are to find the correct sequence of data pre-processing techniques and their use so that the results can be optimal. Topic modeling uses automatic labeling for topic naming. All backend processes are done automatically from the developed system.

ACKNOWLEDGEMENTS

Authors thanks to Scholarship PasTi, a sponsor for master education in the Postgraduate Program of IPB University. Prof. Ir. Teddy Mantoro, MSc., PhD., SMIEEE and Transmissia Semiawan, BSCS., MIT, Ph.D., who are willing to become experts for labeling the results of modeling topics.

REFERENCES

[1] D. a Pendlebury, “Using bibliometrics in evaluating research,” in *Thomson Reuters*, 2008.

[2] A. Öchsner, *Introduction to Scientific Publishing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[3] F. Narin, “Patent bibliometrics,” *Scientometrics*, vol. 30, no. 1, pp. 147–155, May 1994, doi: 10.1007/BF02017219.

[4] Kemenristekdikti, “Technical instructions for regulation of the minister of research, technology and higher education number 20 of 2017 concerning lecturer professional allowances and honorary professor allowances (in *Indonesia*),” 2017. [Online]. Available: https://peraturan.bpk.go.id/Home/Details/140850/permen-ristekdikti-no-20-tahun-2017.

[5] S. Lafia, W. Kuhn, K. Caylor, and L. Hemphill, “Mapping research topics at multiple levels of detail,” *Patterns*, vol. 2, no. 3, Mar. 2021, doi: 10.1016/j.patter.2021.100210.

[6] C. Chaomei, “Science mapping: a systematic review of the literature,” *Journal of Data and Information Science*, vol. 2, no. 3, pp. 1–40, 2017, doi: 10.1515/jdis-2017-0006.

[7] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, “How to conduct a bibliometric analysis: An overview and guidelines,” *Journal of Business Research*, vol. 133, pp. 285–296, Sep. 2021, doi: 10.1016/j.jbusres.2021.04.070.

[8] B. Yoon and Y. Park, “A text-mining-based patent network: Analytical tool for high-technology trend,” *The Journal of High Technology Management Research*, vol. 15, no. 1, pp. 37–50, Feb. 2004, doi: 10.1016/j.hitech.2003.09.003.

[9] T. M. Abuhay, Y. G. Nigatie, and S. V. Kovalchuk, “Towards predicting trend of scientific research topics using topic modeling,” *Procedia Computer Science*, vol. 136, pp. 304–310, 2018, doi: 10.1016/j.procs.2018.08.284.

[10] S. M. Weiss, N. Indurkhya, and T. Zhang, *Fundamentals of predictive text mining*, vol. 44, no. 8. London: Springer London, 2010.

[11] B. V. Looy and T. Magerman, “Using text mining algorithms for patent documents and publications,” in *Springer Handbooks*, 2019, pp. 929–956, doi: 10.1007/978-3-030-02511-3_38.

[12] D. Westergaard, H.-H. Stærfeldt, C. Tønsberg, L. J. Jensen, and S. Brunak, “A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts,” *PLOS Computational Biology*, vol. 14, no. 2, Feb. 2018, doi: 10.1371/journal.pcbi.1005962.

[13] B. B. L. Penning de Vries, M. V. Smeden, F. R. Rosendaal, and R. H. H. Groenwold, “Title, abstract, and keyword searching resulted in poor recovery of articles in systematic reviews of epidemiologic practice,” *Journal of Clinical Epidemiology*, vol. 121, pp. 55–61, May 2020, doi: 10.1016/j.jclinepi.2020.01.009.

[14] M. Kim, Y. Park, and J. Yoon, “Generating patent development maps for technology monitoring using semantic patent-topic analysis,” *Computers & Industrial Engineering*, vol. 98, pp. 289–299, Aug. 2016, doi: 10.1016/j.cie.2016.06.006.

[15] R. Albalawi, T. H. Yeap, and M. Benyoucef, “Using topic modeling methods for short-text data: a comparative analysis,” *Frontiers in Artificial Intelligence*, vol. 3, Jul. 2020, doi: 10.3389/frai.2020.00042.

[16] V. K. Garbhapu, “A comparative analysis of latent semantic analysis and latent dirichlet allocation topic modeling methods using Bible data,” *Indian Journal of Science and Technology*, vol. 13, no. 44, pp. 4474–4482, Nov. 2020, doi: 10.17485/IJST/v13i44.1479.

[17] D. Chehal, P. Gupta, and P. Gulati, “Retracted article: Implementation and comparison of topic modeling techniques based on user reviews in e-commerce recommendations,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 5, pp. 5055–5070, May 2021, doi: 10.1007/s12652-020-01956-6.

[18] J. Choi and Y.-S. Hwang, “Patent keyword network analysis for improving technology development efficiency,” *Technological Forecasting and Social Change*, vol. 83, no. 1, pp. 170–182, Mar. 2014, doi: 10.1016/j.techfore.2013.07.004.

[19] S. Agarwala, A. Anagawadi, and R. M. Reddy Guddeti, “Detecting semantic similarity of documents using natural language processing,” *Procedia Computer Science*, vol. 189, pp. 128–135, 2021, doi: 10.1016/j.procs.2021.05.076.

[20] M. B. Negahban and N. Zarifsanaiey, “Network analysis and scientific mapping of the e-learning literature from 1995 to 2018,” *Knowledge Management & E-Learning: An International Journal*, vol. 12, no. 3, pp. 268–279, Sep. 2020, doi: 10.34105/j.kmel.2020.12.014.

[21] A. R. Hakim, T. Djatna, and A. Febransyah, “Technology map: a text mining and network analysis approach,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 3, no. 1, p. 200, Jun. 2016, doi: 10.11591/ijeecs.v3.i1.pp200-208.

[22] M. Pachayappan and R. Venkatesakumar, “A graph theory based systematic literature network analysis,” *Theoretical Economics Letters*, vol. 08, no. 05, pp. 960–980, 2018, doi: 10.4236/tel.2018.85067.

[23] W. Robinson, A. Dutta, S. Li, and S. Turkelson, “Framework for determining research gaps during systematic review: evaluation,” *Methods Research Report; AHRQ Publication*, vol. 1, no. 3, pp. 1-14, 2013.

[24] I. Badea and S. Trausan-Matu, “Text analysis based on time series,” in *2013 17th International Conference on System Theory, Control and Computing (ICSTCC)*, Oct. 2013, pp. 37–41, doi: 10.1109/ICSTCC.2013.6688932.

[25] L. Martínez-Acosta, J. P. Medrano-Barboza, Á. López-Ramos, J. F. R. López, and Á. A. López-Lambraño, “SARIMA approach to generating synthetic monthly rainfall in the sinú river watershed in Colombia,” *Atmosphere*, vol. 11, no. 6, Jun. 2020, doi: 10.3390/atmos11060602.

[26] S. Bang, R. Bishnoi, A. S. Chauhan, A. K. Dixit, and I. Chawla, “Fuzzy logic based crop yield prediction using temperature and rainfall parameters predicted through ARMA, SARIMA, and ARMAX models,” in *2019 Twelfth International Conference on Contemporary Computing (IC3)*, Aug. 2019, pp. 1–6, doi: 10.1109/IC3.2019.8844901.

[27] K. E. ArunKumar, D. V. Kalaga, C. M. S. Kumar, G. Chilkoor, M. Kawaji, and T. M. Brenza, “Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: auto-regressive integrated moving average (ARIMA) and seasonal auto-regressive integrated moving averag,” *Applied Soft Computing*, vol. 103, May 2021, doi: 10.1016/j.asoc.2021.107161.

[28] K. He, L. Ji, C. W. D. Wu, and K. F. G. Tso, “Using SARIMA–CNN–LSTM approach to forecast daily tourism demand,” *Journal of Hospitality and Tourism Management*, vol. 49, pp. 25–33, Dec. 2021, doi: 10.1016/j.jhtm.2021.08.022.

[29] U. Fayyad and P. Piatetsky-Shapiro, and G. Smyth, “From data mining to knowledge discovery in databases,” *IA Magazine*, vol. 17, no. 3, pp. 37–54, 1993, doi: 10.1609/aimag.v17i3.1230.

[30] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, 1st ed. Cambridge University Press, 2008.

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, 2013.

[32] D. Saxena, S. K. Saritha, and K. N. S. S. Prasad, “Survey paper on feature extraction methods in text categorization,” *International Journal of Computer Applications*, vol. 166, no. 11, pp. 11–17, May 2017, doi: 10.5120/ijca2017914145.

[33] D. Hovy, *Text Analysis in Python for Social Scientists,* Cambridge University Press, 2021.

[34] C. B. G. Salton, “Term-weighting approaches in automatic text retrieval,” *Journal of Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1998.

[35] F. Zhang, H. Fleyeh, X. Wang, and M. Lu, “Construction site accident analysis using text mining and natural language processing techniques,” *Automation in Construction*, vol. 99, pp. 238–248, Mar. 2019, doi: 10.1016/j.autcon.2018.12.016.

[36] G. Liu, M. Boyd, M. Yu, S. Z. Halim, and N. Quddus, “Identifying causality and contributory factors of pipeline incidents by employing natural language processing and text mining techniques,” *Process Safety and Environmental Protection*, vol. 152, pp. 37–46, Aug. 2021, doi: 10.1016/j.psep.2021.05.036.

[37] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, no. 4–5, pp. 993–1022, 2003, doi: 10.1016/b978-0-12-411519-4.00006-9.

[38] D. Blei, L. Carin, and D. Dunson, “Probabilistic topic models,” *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 55–65, Nov. 2010, doi: 10.1109/MSP.2010.938079.

[39] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, Feb. 2015, pp. 399–408, doi: 10.1145/2684822.2685324.

[40] J. Han, M. Kamber, and J. Pei, *Data Mining : Concepts and Techniques : Concepts and Techniques*, 3rd ed. Elsevier, 2012.

[41] S. Park, Y. Yuan, and Y. Choe, “Application of graph theory to mining the similarity of travel trajectories,” *Tourism Management*, vol. 87, p. 104391, Dec. 2021, doi: 10.1016/j.tourman.2021.104391.

[42] J. Hu and Y. Zhang, “Measuring the interdisciplinarity of big data research: a longitudinal study.,” *Online Information Review*, vol. 42, no. 5, pp. 681–696, 2018, doi: 10.1108/OIR-12-2016-0361.

[43] M. E. J. Newman, *Networks*, 2nd ed. Oxford: Oxford University Press, 2018.

[44] K. R. Saoub, “Graph Theory : an introduction to proofs, algorithms, and applications,” in *Textbooks in Mathematics*, 1st ed., London: Chapman and Hall/CRC, 2021.

[45] M. Grandjean, “Introduction to social network analysis : basics and historical specificities,” *HNR+ResHist Conference*, pp. 1–21, 2021, doi: 10.5281/zenodo.5083036.

[46] H. Akaike, “A new look at the statistical model identification,” in *IEEE transactions on automatic control*, 1974, pp. 215–222.

[47] D. C. S. Bisht and M. Ram, *Recent Advances in Time Series Forecasting*. Boca Raton: CRC Press, 2021, doi: 10.1201/9781003102281.

BIOGRAPHIES OF AUTHORS

Asep Herman Suyanto is a student Postgraduate Program at the Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB University, Indonesia. Bachelor's degree from Universitas Gadjah Mada, Indonesia. His research interests include machine learning, data mining, text mining, information retrieval, and graph theory. He can be contacted at email: asep_hs@apps.ipb.ac.id, asep_hs@yahoo.com.

Taufik Djatna is a lecturer and Professor at the Department of Agro-Industrial Technology, Faculty of Agricultural Technology, IPB University, Indonesia. His research interests include blockchain engineering, Kansei engineering, recommendation systems, cyber-physical systems, and customer relationship management. His educational background is a Bachelor's degree from IPB University, Indonesia, a Master's degree from IPB University, Indonesia, Doctoral degree from Hiroshima University, Japan. He can be contacted at email: taufikdjatna@apps.ipb.ac.id.

Sony Hartono Wijaya is a lecturer at the Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB University, Indonesia. His research interests include bioinformatics, machine learning, data mining, information retrieval, and software engineering. His educational background is a Bachelor's degree from IPB University, Indonesia, a Master's degree from University of Indonesia, Indonesia, Doctoral degree from Nara Institute of Science and Technology, Japan. He can be contacted at email: sony@apps.ipb.ac.id.